



*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

**Give to AgEcon Search**

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*



**Proceedings of the 36th Annual Meeting  
Transportation Research Forum**

*Volumes  
1 and 2*

**November 3-5, 1994**

**Daytona Beach, Florida**

**Published and Distributed by:**

**Transportation Research Forum  
1730 North Lynn Street, Suite 502  
Arlington, VA 22209**

# A Multiple Discriminant Analysis of Vessel Accidents

Conway T. Rucks and Louis A. Le Blanc\*

## ABSTRACT

A large sample of 936 vessel accident cases occurring between 1979 and 1987 on the Lower Mississippi River were cluster analyzed to generate four groups relatively unique in their respective attribute values. The attributes used to cluster the accidents included participation in the U. S. Coast Guard's New Orleans Vessel Traffic Service (NOLA-VTS), type of accident, river stage, traffic level, system utilization (total vessel movements which were VTS-supported), accident location, weather conditions, and time of accident. The four-group cluster solution resulted in logical groupings, given the realities of navigating the lower Mississippi River.

The four groups resulting from the cluster analysis were characterized as *Group 1: Danger Zone*, one hundred percent NOLA-VTS participants whose accidents occurred primarily on the most dangerous part of the river; *Group 2: Bad Conditions for Good Navigators*, characterized by a high rate of participation and unserious accidents occurring in treacherous navigating conditions; *Group 3: Probably Preventable*, characterized by low participation rates and serious accidents occurring in not the worst navigating conditions; and *Group 4: Accidents That Shouldn't Have Happened*, characterized by zero participation and serious accidents occurring in reasonable navigating conditions.

In the subsequent discriminant analysis, three discriminant functions correctly classified 96 percent of the total accidents, including 100 percent of Group 1 and Group 4, 90 percent of Group 2, and 88 percent of Group 3. The variables contributing most to overall group differentiation were participation in the system, system utilization, river stage, traffic level, time, and location of the accident. The set of three discriminant functions were statistically significant at a very high level, with each individual function accounting for a large relative percentage of the variance between groups and being correlated with the groups of accidents. In order of discriminating power, the functions could be characterized as (1) "System Participation and Utilization," (2) "Navigating Conditions," and (3) "Time and Place." Thus significant incremental participation rates for the marine tracking technology across the accident clusters effectively distinguishes between casualty groups.

## INTRODUCTION

For a large sample of 936 vessel accident cases that occurred between 1979 and 1987 on the lower Mississippi River, Le Blanc and Rucks (1994) employed cluster analysis to generate four groups relatively unique in their respective attribute values. These categories are logical groups, giving the realities of navigating the lower Mississippi River.

Clustering (Anderberg, 1973; Everitt, 1980) provides an analytical technique which can be used to develop meaningful subgroups of vessel accidents. The groups are not predefined. Cluster analysis classifies a sample of entities into a small number of mutually exclusive groups

based on the similarities among the entities. By studying groups of vessel accidents, research can determine the characteristics or attributes that the casualties share, as well as those in which they differ.

Discriminant analysis is a statistical technique which allows the researcher to study the difference(s) between two or more a priori defined groups of objects with respect to several variables simultaneously. To appropriately use discriminant analysis, the cases should be members of two or more mutually exclusive groups. The data are the values of the variables for cases whose group membership is already known. Researchers also may want to identify the variables that are important for distinguishing among the groups and to develop a procedure for predicting group membership for new cases whose group membership is undetermined. Discriminant analysis is the statistical technique most commonly used to investigate this set of problems. The purpose of this paper is to: 1) determine if statistically significant differences exist between the profiles of the four accidents groups determined by Le Blanc and Rucks (1994) with cluster analysis; and, 2) determine which of the independent variables account most for the differences in the average score profiles of the four accident groups.

## DISCRIMINANT ANALYSIS

The concept underlying discriminant analysis is straightforward. Linear combinations of the independent variables (i.e., predictor variables) are formed and serve as the basis for classifying cases into one of the groups (Fisher, 1936). For the linear discriminant function(s) to be optimal (e.g., to provide a classification rule that minimizes the probability of misclassification), certain assumptions about the data must be met. Each group must be a sample from a multivariate normal population, and the population covariance matrices must all be equal. This permits the precise computation of tests of significance and probabilities of group membership (Blalock, 1979; Lachenbruch, 1975).

"Discriminant analysis" (Hair et al., 1992) is a broad term which refers to several closely related statistical activities, those related to interpreting the group differences and those employed to classify cases into the groups. Interpretation relates to studying the ways in which groups differ: can the researcher "discriminate" among the groups on the basis of some set of characteristics? How well do the characteristics discriminate? And which characteristics are the most powerful discriminators? The other application is to derive one or more mathematical equations for the purpose of classification. These equations, termed "discriminant functions," combine the group characteristics in a way that will allow one to identify the group which a case most closely resembles. Researchers use discriminant analysis for both interpretation and classification.

Discriminant analysis involves deriving the linear combination of the two (or more) independent variables that will discriminate best between the a priori defined groups. This is achieved by the statistical decision rule of maximizing the between-group variance relative to the within-group variance, expressed as the ratio of the between-group to within-group variance. The linear combination for a discriminant analysis are derived from an equation which takes the following form:

$$Z = W_1X_1 + W_2X_2 + W_3X_3 + \dots W_nX_n \quad [1]$$

where:

**Z** = the discriminant score

**W** = the discriminant weights

**X** = the independent variables

Discriminant analysis is the appropriate statistical technique for testing the hypothesis that the group means of the two or more groups are equal. To do so, discriminant analysis multiplies each independent variable by its corresponding weight and adds these products together (see above equation). The result is a single composite discriminant score for each individual cases in the analysis. By averaging the discriminant scores for all the cases within a particular group, the analysis arrives at the group mean. This group mean is referred to as a centroid. When the analysis involves two groups, there are two centroids; with three groups there are three centroids, and so forth. The centroids indicate the most typical location of an individual case from a particular group, and a comparison of the group centroids indicate how far apart the groups are along the dimension being tested.

The test for the statistical significance of the discriminant function is a generalized measure of the distance between the group centroids. It is computed by comparing the distribution of the discriminant scores for the two or more groups. If the overlap in the distribution is small, the discriminant function separates the groups well. If the overlap is large, the function is a poor discriminator between the groups.

A review of the objectives for applying discriminant analysis should further clarify its nature. These include:

- 1) Determining if statistically significant differences exist between the average score profiles of the two (or more) a priori defined groups.
- 2) Establishing procedures for classifying statistical units (cases or records) into groups on the basis of their scores on several variables.
- 3) Determining which of the independent variables account for the differences in the average score profiles of the two or more groups.

As can be noted from the above objectives, discriminant analysis is useful when the analyst is interested either in understanding group differences or in correctly classifying statistical units into groups or classes. Discriminant analysis, therefore, can be considered either a type of profile analysis or an analytical predictive technique. In either case, it is most appropriate where there is a single categorical dependent variable and several metrically scaled independent variables.

## MARINE CASUALTY DATABASE

In this research process, the choice of variables to be used with cluster analysis is the most critical step, i.e., to find that set of variables which best represents the concept of similarity. If important variables are excluded, poor or misleading findings may result. Ideally, variables should be chosen within the context of an explicitly stated theory that is used to support the classification. The theory is the basis for the rational choice of the variables to be used in the study. In practice, however, the theory that supports any classification research is often implicit. For this study, recognized experts in maritime safety for the lower Mississippi River were surveyed to determine the most critical variables to use in this cluster analysis.

The terms "case," "entity," "object" and "pattern" denote the accident data being classified; whereas, "variable," "attribute," "characteristics" and "feature" of the accidents are used to assess the cases' similarity.

The variables for this cluster analysis were identified as the result of discussions with members of the Port Safety Council, Port of New Orleans. Members of the Council include representatives from the following maritime groups: U.S. Coast Guard (Captain of the Port, Marine Inspection Office, Commanding Officer of NOLA-VTS); U.S. Army Corps of Engineers; dock and harbor authorities (Ports of Baton Rouge and New Orleans); steamship companies; barge lines and towing firms; pilot associations; and shippers (Exxon Shipping Co., Texaco, Inc., Chevron U.S.A.). All maritime parties who were interviewed agreed that the following variables concerning vessel safety were most important for any study of accidents on the lower Mississippi River.

The New Orleans Vessel Traffic Service (NOLA-VTS) monitors vessel traffic on the lower Mississippi River from Devil's Swamp near Baton Rouge south to the Gulf of Mexico beyond New Orleans - a distance of about 250 miles. As a voluntary vessel movement reporting system, NOLA-VTS provides ocean-going ships, large barge tows, and other river traffic with pertinent information to aid in making decisions as to the proper navigation strategy in plying the lower Mississippi River. In addition, the NOLA-VTS audits vessel accidents and records them on a daily log.

This database was arranged so as to be consistent with the six operational sectors of the lower Mississippi River as delineated by NOLA-VTS, the primary data source. Table 1 shows the six VTS sectors and their respective mile marks.

The data set for this research study is comprised of 936 cases of vessel accident data occurring between 1979 and 1987 on the lower Mississippi River. Each accident record consists of the following attributes: 1) whether or not the vessel(s) involved in an accident were participating in the vessel tracking; 2) accident type; 3) river stage; 4) traffic level; 5) utilization rate of the U.S. Coast Guard's vessel tracking service; 6) accident location; 7) weather; and, 8) time of the accident. Table 1A and Table 1B provide descriptive statistics about the following variables used in the cluster analysis of vessel accidents.

TABLE 1A

## Descriptive Statistics - Continuous Variables

	Frequency	Percent	Mean	Min	Max
<b><u>River Stage</u></b>					
.1 - 4.5	309	33.0	9.2	.1	17.2
4.5 - 9.2	180	19.2			
9.2 - 13.2	128	13.7			
13.2 - 17.2	319	34.1			
<b><u>Traffic Level</u></b>					
3 - 130	6	.6	256	3	466
130 - 256	474	50.6			
256 - 361	373	39.9			
361 - 466	82	8.8			
<b><u>Utilization</u></b>					
3.0 - 18.4	7	.7	33.8	3.0	65.0
18.4 - 33.9	469	50.1			
33.9 - 49.4	449	48.0			
49.4 - 65.0	11	1.2			

**Participation**

This dichotomous variable indicates whether or not a vessel, which was involved in an accident, was participating in the vessel tracking service. This indicates communication between vessel and NOLA-VTS's control center and the exchange of pertinent information to aid navigation efforts.

**Accident Type**

According to NOLA-VTS, casualty types include: 1) collisions; 2) rammings; and, 3) groundings. A collision is defined as any contact between vessels which are underway, anchored, moored, or in the process of docking or undocking. A ramming is the collision of a vessel with a fixed object such as a wharf, dock, pier, bridge, submerged object, or aid to navigation. Groundings represent vessel contact with the river bottom, and may or may not result in damage to the vessel.

TABLE 1B

## Descriptive Statistics - Categorical Variables

	Frequency	Percent
<b><u>Accident Type</u></b>		
Collision	207	22.1
Ramming	422	45.1
Grounding	297	31.7
<b><u>Time</u></b>		
0:00 - 5:59 AM	266	28.4
6:00 -11:59 AM	178	19.0
12:00 - 17:59 PM	230	24.6
18:00 - 23:59 PM	262	28.0
<b><u>Participation</u></b>		
Non-Participant	505	54.0
Participant	428	45.7
<b><u>Weather</u></b>		
Dec - Jan	172	18.4
Feb - Mar	120	12.8
Apr - May	295	31.5
Oct - Nov	92	9.8
Jun - Sep	257	27.5
<b><u>Location</u></b>		
I - Miss River	194	20.7
IA - Gulf Outlet	14	1.5
II - Miss River	228	24.4
IIA - Gulf Outlet	9	1.0
III - Miss River	169	18.1
IV - Miss River	322	34.4

**River Stage**

The stage of the river (height above sea level measured in feet) is a very critical element of vessel safety, because river stage directly determines the velocity of the current. The hazardous conditions that often accompany the changes in the river stage (hence, river current) precipitate many vessel casualties on the lower Mississippi River between Baton Rouge and the Gulf of Mexico. River stage is a continuous variable.



## **Traffic Level**

According to maritime authorities, traffic level is another major element influencing the occurrence of accidents in the study area. The measure of traffic includes ocean- going ships, barge tow assemblies, tugs, as well as excursion craft. Traffic level, as measured by the number of vessel movements, represents the number of vessel transits on the day of the accident. Traffic level is a continuous variable.

## **System Utilization**

The system utilization variable is defined as the percentage of total vessel movements which were VTS-supported on the day of the accident. The U.S. Coast Guard calculates the level of utilization for the NOLA-VTS at the hour of peak daily traffic. System utilization is a continuous variable.

## **Location**

Indicator variables provide the area of the lower Mississippi River where the accident happened. Six areas, each with a different physical geography, were employed. (Refer to Table 1b for frequencies of where accidents occurred.) The percentages of accidents occurring in each location category were used as indicators of the relative strength of the areas in determining accidents.

## **Weather**

Meteorological conditions that are hazardous to navigation in the study area are those which produce strong winds, heavy rainfall, and fog. Indicator variables were constructed to model these seasonal weather patterns which affect navigational safety on the lower Mississippi River (i.e., Dec-Jan, Feb-Mar, Apr-May, Oct-Nov, Jun-Sep). The percentages of accidents occurring in each weather category were used as indicators of the relative strength of the weather groupings in determining accidents. The percentages of accidents occurring in each time category were used as indicators of the relative strength of the time groupings in determining accidents.

## **Time**

The time that a vessel casualty occurs is recorded by the U.S. Coast Guard using the standard military clock. For purposes of this analysis, the time of an accident was indicated by

one of four categories: between midnight and 6:00 AM, between 6:00 AM and noon; between noon and 6:00 PM, and between 6:00 PM and midnight.

Following an initial cluster analysis, the percentage of cases currently classified was tentatively tested by discriminant analysis for each cluster solution from twelve to two groups. The percentage of cases currently classified was highest (96%) for the four-group solution. The groups identified are described below. They are characterized according to differences in their accident characteristics in Table 2.

TABLE 2

Description of Accident Clusters				
	Group 1	Group 2	Group 3	Group 4
Cases	224	226	133	340
V1	1.00000	.67699	.36842	.00000
V2	62% Ram <sup>a</sup> 86% C&R <sup>b</sup>	66% Ground <sup>c</sup>	53% Ram 73% C&R	49% Ram 77% C&R
V3	Slow Current <sup>d</sup> 69% < 7 Feet	Fast Current 73% > 14 Feet	Slow Current 58% < 7 Feet	Slow Current 48% < 7 Feet
V4	Moderate	Heavy	Moderately Heavy	Moderate
V5	Moderate	Heavy	Moderately Heavy	Moderate
V6	Concentrated Sector IV	Scattered	Scattered	Scattered
V7	36% Bad Weather	74% Apr-May Bad Weather	43% Bad Weather	45% Bad Weather
V8	Dispersed	Dispersed	83% Mornings 94% Daylight	0 Mornings 73% Night

<sup>a</sup>Ram is abbreviation for ramming.

<sup>b</sup>C&R represents collisions and rammings combined.

<sup>c</sup>Ground refers to a grounding of a vessel on the river bottom.

<sup>d</sup>Speed of current is directly related to river stage (i.e., the lower the river stage the slower the current and vice versa.

V1 = Participation; V2 = Accident Type; V3 = River Stage; V4 = Traffic;

V5 = Utilization; V6 = Location; V7 = Weather; and, V8 = Time.

## DESCRIPTION OF ACCIDENT GROUPS

The following discussion of the accident groups emphasizes the apparent general impact of the vessel tracking service (VTS) participation rate across the four groups. Despite their obvious influence, the other seven attributes of the accident cases had apparently varying effects on particular clusters.

### Group 1 - *Danger Zone*

This cluster is a very interesting and unique group, especially in reference to group 4. Group 1 may be characterized as *Danger Zone*. Group 1 experienced a 100 percent participation rate, while group 4 had zero participation. Group 1 contains about 25 percent of the accident cases. As for attribute values in cluster 1, ramblings comprise 62 percent of accident types, while collisions add another 24 percent. Together, these potentially serious casualties total 86 percent of accident types. There is no particular time segment when these accidents are more likely to occur. Thirty-six percent of the casualties happen during bad weather, slow current, moderate traffic levels, and moderate (i.e., not low) utilization levels of the VTS along the lower Mississippi River.

However, the location of the collisions and ramblings is a very distinctive feature of this cluster. Fifty-six (56) percent of the accident cases are located in Sector IV of the lower Mississippi River. This section of river is especially dangerous for vessel traffic because of the myriad of narrow twisting turns, making navigation very hazardous for large ships and barge tows which are destined for the heavy concentration of petrochemical refineries and grain elevators in this area of the Mississippi River. This group can be characterized by river mariners attempting to provide safe passage for their vessels, cargoes and crews in treacherous waters as evinced by the 100 percent participation rate, but experiencing more than half of their accidents (typically collisions and ramblings) in a most dangerous sector of the lower Mississippi River.

### Group 2 - *Bad Conditions for Good Navigators*

Group 2 may be characterized as *Bad Conditions for Good Navigators*. This cluster represents a dramatically different set of attribute values than does group 4. In general, group 2 is characterized by relatively poor circumstances for navigation, such as fast current, heavy traffic and the bad weather conditions of April and May (i.e., 73 percent of these cases occurred in these two months). The location and time attributes were not concentrated in any particular category but were distributed across these respective dimensions. The distinguishing attribute values for this group were: the high utilization levels (i.e., 80 percent of the cases were at "high" utilization rates) of the tracking service by mariners across the system at the time of these accidents; the high participation rate (67 percent) in the VTS for these accident cases in group 2; and that the majority (66 percent) of accident types was a grounding.

A grounding is not a serious marine casualty like a collision or ramming. A grounding is defined as the vessel touching the river bottom, without regard to damage to the craft or whether the vessel was immobilized. Despite poor or even treacherous conditions, these accident cases in group 2 were most likely to be a non-serious casualty. Coupled with the high participation rate, these attribute values indicate that these mariners were doing a good job during the most trying of conditions, in contrast to the accident cases in group 4.

### **Group 3 - *Probably Preventable***

Group 3 may be characterized as *Probably Preventable*. This cluster exhibits attribute values that lie between the initial two groups already discussed. These intermediate characteristics depict mediocre conditions for navigation and associated accidents. This group of entities, with the least number of assigned cases (e.g., 133 entities or 14 percent), manifested slow river current and moderately heavy traffic. System-wide utilization of the NOLA-VTS was moderately heavy, while the location of accidents was scattered along the lower Mississippi River and not concentrated in any particular sector. About 43 percent of these accidents occurred during the months with bad weather for navigation. Of particular interest, 83 percent of these accidents materialized during the morning hours between 6:00 AM and noon, and 94 percent of all accidents happened during the usual daylight hours (e.g., 6:00 AM to 6:00 PM). Collisions and rammings accounted for 73 percent of all accidents in group 3, while the participation rate was about 37 percent.

Group 3 is relatively similar to group 4. A distinct property of both these groups was a relatively low (39 percent) in the former and the non-existent participation rate in the latter for NOLA-VTS at the moment of the incident. The serious nature of collisions and rammings, combined with moderate or even higher levels of VTS utilization as well as tolerable maritime conditions, suggests similar but nevertheless different groups. Group 3 is distinct because of the higher participation feature and the very high percentage of accidents that occurred during the morning hours. Group 4 experienced 73 percent accidents at night.

### **Group 4 - *Accidents that Shouldn't Have Happened***

Group 4 may be characterized as *Accidents that Shouldn't Have Happened*. This cluster consists of 340 accident cases, the largest cluster containing more than one-third of the sample of 923. By inspection of the descriptive information in Table 2, the attribute values of this group describe generally good navigational conditions, such as slow current, moderate traffic levels, and moderate (i.e., not low) utilization levels of the VTS along the lower Mississippi River. Other characteristics of these entities include a preponderance of collisions and rammings which usually occurred at night and not concentrated in any particular sector of the river. Forty-five percent of accidents occurred during months enduring bad weather (i.e., December, January, April and May), indicating that bad weather as expected contributed to casualties (Le Blanc and Kozar, 1990).

The most distinctive property of this group was that *none* of these vessels involved in an accident was participating in the NOLA-VTS at the moment of the incident. This phenomena, added to the serious nature of collisions and ramblings along with moderate levels of VTS utilization by mariners plying the river at the time of the accident, indicates a powerful influence to ascribe these accident cases to cluster 4. These accident cases reveal potentially serious casualties, despite reasonable navigational conditions and in conjunction with no mariner in this group partaking of the tracking service.

As indicated above, the four-group cluster solution was selected as optimum using the criterion of the percentage of cases currently classified into groups via discriminant analysis. From two to twelve-group cluster solutions, the percentage of cases correctly classified was highest (96%) for the four-group solution. A discussion and analysis of the four-group solution is presented following.

## FOUR-GROUP DISCRIMINANT ANALYSIS

### The Number of Functions

When there are four groups, three discriminant functions can be calculated (assuming there are three or more predictors). The first function has the largest ratio of between-groups to within-groups sums of squares. The second function is uncorrelated with the first and has the next largest ratio. In general, if there are  $k$  groups,  $k - 1$  discriminant functions can be computed. They are all uncorrelated with each other and maximize the ratio of between-groups to within-groups sums of squares, subject to the constraint of being uncorrelated.

### The Raw Discriminant Function Coefficients

Table 3 contains the three sets of unstandardized (raw) discriminant function coefficients for the vessel accident cases. Based on these coefficients, it is possible to computer three scores for each case, one for each function. When there are several groups (four), a case's values on all functions (three) must be considered simultaneously in determining to which group a particular case will be assigned. These raw coefficients are totally uninterpretable as coefficients, and the scores they produce for the accident cases have no intrinsic meaning.

### Classification Results

Table 4, sometimes termed the *confusion matrix*, shows the numbers of correct and incorrect classifications for each group. Only the cases with complete information for all predictor variables are included in the classification results table. Correctly classified cases appear on the diagonal of the table.

TABLE 3

## Unstandardized Discriminant Coefficients

	Function 1	Function 2	Function 3
Participation	2.1799060	-1.9686060	.3473889
Accident Type	.2489460	.4486176	-.0553743
River Stage	.0693566	.1123099	.0066109
Traffic Level	.0080448	.0087239	-.0007281
Utilization	.0337605	.0245997	-.0149794
Location	.0287266	-.0124182	-.0745309
Weather	.0399863	.0351888	-.0066928
Time	-.0466541	.0736156	.3582734
Constant	-6.0813310	-6.5032120	-10.3063100

TABLE 4

## Classification Results

Cluster	Number of Cases	Predicted Group Membership			
		1	2	3	4
Group 1	224	224 100%	0 0%	0 0%	0 0%
Group 2	233	9 3.9%	209 89.7%	2 0.9%	13 5.6%
Group 3	134	10 7.5%	0 0%	118 88.1%	6 4.5%
Group 4	345	0 0%	0 0%	0 0%	345 100%

Percent of "grouped" cases correctly classified: 95.73%

For group 1 and group 4, one-hundred percent of their respective accident cases was correctly classified. For group 2, 89.7 percent were correctly classified, or 209 of 233 cases. The discriminant functions correctly classified 88.1 percent of the vessel casualty cases in group 3. Overall, the percent of "grouped" accident cases correctly classified was 95.73 percent.

## The Standardized Coefficients

Once the discriminant functions have been derived, their meanings can be interpreted by studying the individual variables and the functions. When more than one function exists, the researcher should question whether all of them are needed.

To discover the contributions of the individual variables, the analysis must go beyond the unstandardized coefficients. While the unstandardized coefficients do tell us the absolute contribution of a variable in determining the discriminant score, this information may be misleading when the meaning of one unit change in the value of a variable is not the same from one variable to another (i.e., when the standard deviations are not the same). If the researcher wants to know the relative importance of the variables, they must look at standardized coefficients, which are the ones that would be obtained if the original data all had standard deviations of one. This can be achieved by converting the raw data into standard form (i.e., values on a given variable have been adjusted so that they have a mean of zero and a standard deviation of one).

TABLE 5

### Standardized Discriminant Coefficients

	Function 1	Function 2	Function 3
Participation	.80434	-.58216	.10273
Accident Type	.17104	.30822	-.03804
River Stage	.32769	.53063	.03123
Traffic Level	.46902	.50881	-.04247
Utilization	.21565	.15714	-.09568
Location	.20414	-.08825	-.52964
Weather	.27342	.24062	-.04576
Time	-.12504	.19730	.96021

Table 5 reports the standardized discriminant coefficients for the three functions. For function 1, *Participation* (.80) makes the greatest contribution, and *Traffic Level* (.47) and *River Stage* (.38) to a lesser extent. The five other variables in Function 1 are of relatively less importance, contributing at most a third as much as *Participation*. On Function 2, three of the variables (*Participation*, *River Stage*, and *Traffic Level*) have relatively high standardized coefficients. So each makes a somewhat similar contribution to the discriminant function of this dimension. *Time* followed by *Location* are the dominant variables on the third function.

### Within-Groups Structure Coefficients

To determine the similarity between a single variable and a discriminant function, the analyst can look at the correlation between the two. A structure coefficient indicates how closely a variable and a function are related. When the absolute magnitude of the coefficient is very large (near +1.0 or -1.0), the function is carrying nearly the same information as the variable. When the coefficient is near zero, they have very little in common. The researcher can "name" a function on the basis of the structure coefficients by noting the variables having the highest values. If those variables seem to be measuring a similar characteristic, the function could be named after that characteristic.

Structure coefficients yield information quite different from what is communicated by the standardized coefficients. The standardized coefficients give the variable's contribution to calculating the discriminant score. This is but one way of looking at the variable's importance. If two variable share nearly the same discriminating information (i.e., if they are highly correlated), they must share their contribution to the score. Consequently, their standardized coefficients may be smaller than when only one of the variables is used. Or the standardized coefficients might be larger but with opposite signs, so that the contribution of one is partially canceled by the opposite contribution of the other. This is because the standardized coefficients take into consideration the simultaneous contributions of all the other variables. The structure coefficients, however, are simple bivariate correlations, so they are not affected by relationships with other variables.

TABLE 6

Pooled Within-Group Correlation Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1	1.000							
V2	.046	1.000						
V3	-.077	-.158	1.000					
V4	.004	.106	-.229	1.000				
V5	-.031	-.063	.006	.003	1.000			
V6	-.144	-.029	.064	-.100	-.036	1.000		
V7	-.029	-.105	.226	-.494	-.018	.008	1.000	
V8	.107	-.031	-.049	.045	.019	-.231	.021	1.000



As shown in Table 6, the pooled within-groups correlation matrix indicates if there is any substantial association between the various attributes. The highest correlation coefficient is  $-.229$  between river stage (the proxy for speed of current) and traffic. This maximum correlation level indicates a minimum association between two variables and suggests that these attributes are very independent features of the accident cases.

TABLE 7

## Within-Groups Structure Coefficients

	General Utilization Function 1	Navigating Conditions Function 2	Time/Place Function 3
Participation	.73108*	-.58485	.12919
Utilization	.16815*	.16188	-.09621
River Stage	.21365	.45016*	.01406
Traffic Level	.37585	.42398*	-.06052
Weather	.27772	.32119*	-.01004
Accident Type	.16231	.21247*	-.07662
Time	-.07633	.15051	.84388*
Location	.07996	-.07972	.30368*

Note: Variables ordered by size of correlation within function. Asterisk indicates variables with large coefficients for a particular function.

Table 7 provides the pooled-within-groups correlations between the three discriminant functions and the variables within the groups. The three functions are named "General Utilization," "Navigating Conditions" and "Time/Place," respectively because of the variables within each function having the largest coefficients. Within Table 7, the variables are listed in order of the size of correlation within the function. (This listing is different than the other two tables with unstandardized and standardized coefficients.) By reading across the row associated with a particular variables, the variables are ordered by the size of their correlation coefficient.

For example by reading across the *Participation* row, Function 1 has the highest correlation coefficient for this variable (.73108) and this is noted by an asterisk. For *Utilization*, Function 1 also has the highest correlation (.16815) and is designated as such with an asterisk. Function 1 has the largest coefficients for these two variables, which measure individual participation and system-wide utilization of the NOLA-VTS. Hence, Function 1 can be named "General Utilization" since it contains the largest coefficients for these two similar measures of information system utilization.

Similarly, Function 2 can be named "Navigating Conditions," as it has the largest coefficients for *River Stage*, *Traffic Level*, *Weather*, and *Accident Type*. These variables are all similar in that they are associated with environmental conditions for navigating the lower Mississippi River. Function 3 is termed "Time/Place," as it has the largest coefficients for *Time* and *Location* of vessel accidents. These temporal and spatial measures of casualties are similar characteristics.

### How Many Functions Are Significant?

**Eigenvalues** - When more than one discriminant function is derived, several statistics are available to determine the effectiveness of the discriminant functions. For each function, the *eigenvalue* is the ratio of between- groups to within-groups sums of squares. Large eigenvalues are associated with "good" functions. Thus, the function with the largest eigenvalue is the most powerful discriminator, while the function with the smallest eigenvalue is the weakest.

TABLE 8

#### Eigenvalues and Measures of Importance

Function	Eigenvalue	Percent Variance	Cumulative Percent	Canonical Correlation
1	2.53786	51.64	51.64	.8469610
2	1.40794	28.65	80.28	.7646620
3	.96917	19.72	100.0	.7015494

Table 8 gives these results. There are three nonzero eigenvalues, and they have been presented in the order of descending magnitude. The size of the eigenvalue is related to the discriminating power of the function. Therefore, Function 1 is the most powerful; the second function provides the greatest discrimination after the first has done its best; the third provides the greatest further discrimination after the first and second have done their best. Taken together, all the functions do not necessarily provide perfect discrimination, but at least the order of importance is known.

The actual numbers representing the eigenvalues are not of any immediate value. They cannot be interpreted directly. When there is more than one function, the researcher may want to compare the relative magnitudes to see how much of the total discriminating power each has. Thus, 2.53786 for the first eigenvalue is only 1.8 times larger than the eigenvalue for the second function. And the fact that the first eigenvalue is only 2.62 times bigger than the third indicates that the third function may be relatively influential.

**Percent Variance** - To assist in this comparison, the eigenvalues can be converted into relative percentages. This is done by summing all the eigenvalues to get a measure of the total discriminating power, then, dividing this sum into each individual eigenvalue. (See Table 8.) Thus Function 1 contains 51.64 percent of the total discriminating power in this system of equation, while Function 2 and Function 3 provide 28.65 percent and 19.72 percent of the discriminating power, respectively.

There is no rule stating how large the relative percentage must be before the function is of interest. However, Substantive significance is the ability of a research finding to have meaning in explaining the phenomenon under investigation. All that the relative percentage can communicate is whether a function is so weak relative to others that it is unlikely to add further to the understanding of the differences between the groups. In this regard, all three discriminant functions in this problem appear to distinguish between group membership. Function 1 accounts for over 50 percent of the variance, while Function 2 and Function 3 explain nearly 30 and 20 percent respectively.

**Canonical Correlation** - Another way to judge the substantive utility of a discriminant function is by examining the canonical correlation coefficient. This coefficient is a measure of association which summarizes the degree of relatedness between the groups and the discriminant function (Levine, 1977). A value of zero denotes no relationship at all, while large numbers (always positive) represent increasing degrees of association with 1.0 being the maximum.

On the basis of Table 8, one should not form the conclusion that the first discriminant function will always have a large canonical correlation. Even though the first function is always the "most" powerful in a relative sense (as measure by the relative percentage), it may be only weakly related to the groups (as measure by the canonical correlations). For this reason, the canonical correlation is more useful, because it reports how well the discriminant function is doing. If the groups are not very different on the variables being analyzed, then all of the correlations will be low, because the analyst cannot create discrimination when none already existed. By examining both the relative percentage and the canonical correlation, we can determine fairly well how many discriminant functions are substantively meaningful and how much utility they have in explaining group differences.

As shown in Table 8, the canonical correlations range from approximately .85 to .70, a spread of only .15 at high levels of association. This indicates that each discriminant function is strongly correlated with the groups of accidents. Combined with the relative percentage of explained variance, each of these functions are very meaningful and provide much utility in clarification of group differences.

### Testing the Significance of the Discriminant Function

To this point, the analysis has discussed the number of discriminant functions in terms of some mathematical limits and substantive importance. These considerations apply without regard to the sampling properties of the data. They are equally appropriate with population data as well as any type of sample. If the analysis is about population data, then the questions of the number of functions and their importance have been settled by the relative percentage and the

canonical correlation. Within the limits of measurement error, these statistics completely describe the degree of discrimination between the groups and the discriminating variables.

When the data are from a sample such as these vessel accidents, as opposed to constituting the entire population, an additional question must be raised. What is the probability that the sampling process produced cases which show the computed degree of discrimination when in fact there are no group differences in the population?

When there are no differences among the populations from which the samples are selected, the discriminant functions reflect only sampling variability. A test of the null hypothesis that the means of all discriminant functions in all groups in the population are really equal and zero can be based on Wilk's lambda. Since several functions must be considered simultaneously, Wilk's lambda is not just the ratio of the between-groups to within-groups sums of squares but is the product of the univariate Wilk's lambda for each function.

The most common test for the statistical significance of the discriminant functions proceeds indirectly. Rather than testing the function itself, the residual discrimination in the system is examined prior to deriving that function. "Residual discrimination" means the ability of the variables to discriminate among the groups beyond the information that has been extracted by the previously computed functions. If the residual discrimination is too small, then it is meaningless to derive any more functions, even if they exist mathematically.

Values of lambda which are near zero denote high discrimination (i.e., the group centroids are greatly separated and very distinct relative to the dispersion within the groups). As lambda increases toward its maximum value of 1.0, it is reporting progressively less discrimination. When lambda equals 1.0, the group centroids are identical (no group differences).

TABLE 9

Residual Discrimination and Test of Significance				
Functions Derived, k	Wilk's Lambda	Chi-Squared	Degrees Freedom	Level of Significance
0	.0596115	2583.00	24	.0000
1	.2108971	1425.60	14	.0000
2	.5078285	620.69	6	.0000

The four groups of vessel accidents are very different on the selected variables, so it would be reasonable to derive discriminant function. After the first (and most powerful) function has been derived, it has removed a good deal of the discriminating information from the model. We now want to inquire whether enough residual discrimination remains to justify the derivation of the second function. From Table 9, Wilk's lambda now equals .2108971 (for  $k = 1$ ), which is still small. Removing the second function depletes the discriminating information further, but lambda becomes .5078285 for  $k = 2$ . This value is not very high and actually in the middle of the range of possible values for lambda. This lambda for  $k = 2$

indicates that the remaining information about group differences may be worth pursuing and is consistent with the indications given by the relative percentages and the canonical correlations. In other words, the third discriminant function is likely important.

Wilk's lambda is another measure of association. But its inverse denotation and its emphasis on the residual discrimination make it less useful than the relative percentage in this regard than the relative percentage and canonical correlation. Lambda, however, can be converted into a test of significance. Thus, it can be employed as an intermediate statistic rather than the desired end product.

The analysis tests the significance of lambda by converting it into an approximation of the chi-square distribution. Table 9 gives the chi-square results for the vessel accident data. The group differences are definitely significant before the derivation of all three discriminant functions. The significance level of .0000 indicates that one would get a chi-square this large or larger zero times out of 10,000 samples when there actually were no differences between the centroids. Given this unlikely event, we are safe in assuming that the results did come from a population which did have differences between the groups. It also indicates that all the derived discriminant functions are statistically significant as a set. It does not indicate the significance of any single function (unless only one has been derived), but rather it gives the significance of all derived functions working together.

## SUMMARY AND CONCLUSIONS

For this large sample of accident cases, cluster analysis generated four logical groupings that are relatively unique in their respective attribute values. Given these groups, multiple discriminant analysis allowed for the interpretation of how these groups differ on the basis of a set of eight attributes. The analysis further determined how well the individual functions discriminated, and which attributes are the most powerful discriminators.

The three discriminant functions correctly classified 95.73 percent of the 936 accident cases. Accident cases were perfectly classified (100 percent correct) for group 1 and group 4. The discriminant functions correctly classified 89.7 of the cases correctly for group 2 and 88.1 percent of the cases in group 3. The set of three discriminant functions were statistically significant at a very high level, with each individual function accounting for a large relative percentage of the variance between groups and being highly correlated with the groups of accidents. The most powerful first function was named "General Utilization," as the *Participation* and *Utilization* attributes had the most powerful coefficients based on their relationship with Function 1. Not surprisingly for group 1 (Danger Zone) and group 4 (Accidents that Shouldn't Have Happened) where all cases were correctly classified, all mariners involved in an accident were characterized as either fully participating (group 1) or not participating at all (group 4) in the computerized vessel tracking service. In other words, the most powerful function 1, where *Participation* and *Utilization* were dominant attributes, likely contributed significantly to the 100 percent correctly classified cases for group 1 and group 4.

In a similar manner, function 2 and function 3 (respectively named "Navigating Conditions" and "Time/Place") likely contributed to the discriminating power between group 2 (Bad Conditions for Good Navigators) and group 3 (Probably Preventable). These

groups were characterized by differences in navigating conditions (i.e., weather, traffic and current) as well as temporal and spatial elements. The individual participation rate of the VTS measures the linkage between specific mariners, the vessel tracking technology, and the system performance dimension of accidents. Accidents serve as an operational measure of marine safety, and specifically the safety of vessels, crews and cargoes. Significant incremental participation rates for the marine tracking technology across the accident clusters effectively distinguishes between casualty groups. From a risk and insurance perspective, these findings might provide justification to require participation in the vessel tracking service, either by governmental agencies or private insurers.

## REFERENCES

- Anderberg, M. R., *Cluster Analysis for Applications*, New York, NY: Academic Press, 1973.
- Blalock, H. M., *Social Statistics*, New York: McGraw-Hill, 1979.
- Everitt, B., *Cluster Analysis*, New York, NY: Halsted Press, 1980.
- Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 1936, 7, 179-188.
- Gilbert, Ethel S., "On Discrimination Using Qualitative Variables," *Journal of the American Statistical Association*, 1968, 63, 1399-1412.
- Hair, J. F., Jr., R. D. Anderson, R. L. Tatham, and W. C. Black, *Multivariate Data Analysis with Readings* (Third Edition), Macmillan Publishing Company, New York, 1992.
- Lachenbruch, P. A., *Discriminant Analysis*, New York: Hafner, 1977.
- Le Blanc, L.A. and Conway T. Rucks, A Cluster Analysis of Vessel Accidents, *Proceedings of the 29th Annual Conference*, Canadian Transportation Research Forum, May 15-18, 1994. Victoria, British Columbia, Canada, pp. 153-165.
- Le Blanc, L. A. and K. A. Kozar, "An Empirical Investigation of the Relationship Between DSS Usage and System Performance: A Case Study of A Navigation Support System," *MIS Quarterly*, 1990, 14, 263-277.
- Levine, M. S., *Canonical Analysis and Factor Comparison*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-006, Beverly Hills and London, Sage Publications, 1977.

**ENDNOTE**

- \* The authors are Professors in the Department of Marketing and Management, respectively, the University of Arkansas.